ISSN No.: 2454- 2024 (online)

International Journal of Technical Research & Science

# A STUDY OF IDS TECHNIQUE USING DATA MINING

Seema Ranga[1], Ajay Jangra[2]

Computer Science Engineering, UIET kurukshetra, Haryana, India

**Abstract-** in the world of communication, Most of our crucial data is stored in a computer remote and in the most cases we exchange it over a network hence security is a big concern. But it's not just our data transmitting over the network but different types of attacks that can harm our stored data. So Monitoring computer system, its logs (administration logs, security logs, system logs, network logs) and protecting our crucial data is necessary. An intrusion detection system is an application that provides protection from malicious activities or policy violations and generates various rules to defend computer security and this system is relevant for intrusion detection [1].

**Index terms:** ids, Security, problems, KDD cup, data mining techniques, k-means clustering.

## 1. INTRODUCTION

The aim of Intrusion detection System is to defend the security of the Computer system by a layer over the defense system. IDS systems sense the misuse, breach in the security system and also the malicious or unauthorized access to the system [1]. Although Firewalls works for the same reason but the major difference between firewalls and the IDS is IDS suspect the source of the attack and signals the alarm to the system but a firewall directly stops the communication without informing the system. These attacks requires true concerns as they harm the data stored in system and also effect the network traffic, data packet etc.

### 1.1 Function of Intrusion Detection Systems

Intrusion detection system performs many functions which are vital for the system. These are as follows: Monitoring and analyzing together the activities of user and system.

- Analyzing system configuration and vulnerabilities
- Assessing system and file integrity.
- Ability to recognize patterns representative of attack.
- Analysis of abnormal activity patterns.
- Tracking user policy violations.

## 2. PROBLEM IN IDS

There are some classes of problems also in intrusion detection system these are discussed below:

### 2.1 Threshold Detection

Positive attributes of user and system behavior are expressed in terms of count with some level established as permissible. Such behavior attributes can include the number of files accessed by user in a given period of time the number of failed attempts to login to the system the amount of CPU utilized by a method. Use this method in Anomaly Based Intrusion Detection System generate a high level of false positives alarms.

### 2.2 False Positives

A false positive occurs when normal attack is incorrectly classified as malicious and treated therefore. The solution is to examine and review the IDS configuration to prevent false positive from occurring again.

### 2.3 Updates LAG:

The main matter occurs to Signature-Based Intrusion Detection System is the update lag. In other words, will be always a lag between the appearances of new thread and the IDS's updates.

### 2.4 False Negatives

A false negative occurs when an attack or an event is either not detected by the IDS or is considered kind by the analyst. Ordinarily the term false negative would only apply to the IDS not coverage an event.

### 2.4 Data Size

The amount of data the analyst can efficiently analyze.

## 3. SECURITUES OF IDS

Ids provide securities for the relevant system these securities are

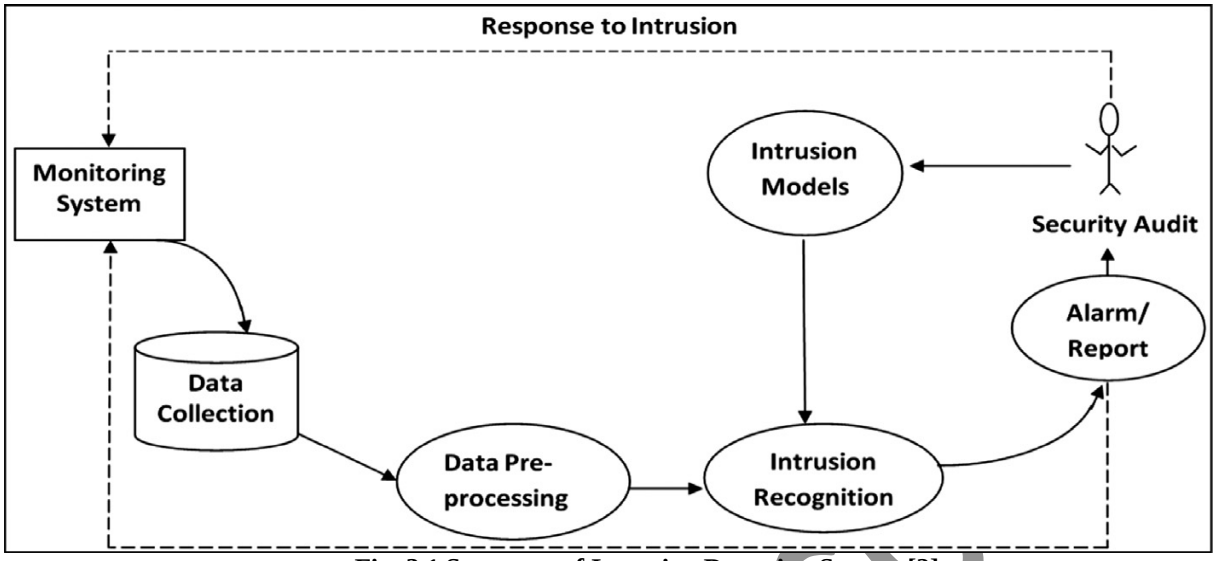


**Fig. 3.1 Structure of Intrusion Detection System [2]**

### 3.1 Data Confidentiality

It checks whether the information stored on the system is protected beside unauthorized access. Since systems are sometimes used to manage sensitive information, data privacy is often a gauge of the ability of the system to protect its data.

### 3.2 Availability

The network should be rough to Denial of Service attacks. Intrusion detection system based on sources of examination information it can be divided into 3 subcategories.

### 3.3 DATA INERGRITY

It refers to maintaining and assuring the correctness and consistency of data over it entire life-cycle. No corruption or data loss is acknowledged either from random events or malicious movement.

## 4. INTRUSION DETECTION SYSTEM BASED ON DATA MINING

Using data mining into intrusion detection system improves the performance has become one of the major concern in the research of intrusion detection. Data mining normally refers to the process of extracting expressive models from large stores of data. The recent rapid development in data mining has made existing a multiplicity of algorithm, drawn from the fields of statistics, pattern recognition, machine learning. Several types of algorithms are mainly useful for mining data. Intrusion detection has great useful drive and application value therefore intrusion detection based on data mining cannot stop at the theoretical investigation. The joint use of several data mining methods can effectively improve data processing speed and quality. Data mining provides decision support for intrusion management. It also discovered unknown patterns of attack or intrusions. In this way it helps intrusion detection organization for detecting new vulnerabilities and intrusions. Data mining can improve deviation detection rate, control false alarm rate and reduce false dismissals. There are many data mining methods. Data mining can be separated into four types: association analysis, sequence analysis, classification analysis and cluster analysis [5]. Association provides simple but valuable description form for the rule mode in data mining i.e. describe invasion of performance patterns. Classification maps the data item into one of the several predefined classes. Data mining is a process of analyzing data from unlike sources and short and snappy it into useful information. It is a process of converting data into [6] Information. Data are facts and information is processed data. Data mining is process of finding correlations or patterns among a large dataset. The proposed mechanism uses data mining algorithms for classification of dataset.

- It can process large amount of data.
- It doesn't need the users' subjective evaluation, and is more likely to determine the ignored and hidden information.
- Those two are especially applicable to the intrusion detection based on analyzing the irregularity of auditing record.

- In order to determine how data mining techniques (DMT) and their applications have developed, during the past decade, this data mining techniques and their applications and development, through a survey of literature and the classification of articles, from 2000 to 2011[16].
- Data mining techniques (DMT) have formed a branch of applied artificial intelligence (AI), since the 1960s. During the intervening decades, important innovations in computer systems have led to the introduction of new technologies [18], for web-based education. Data mining allows a search, for valuable information, in large volumes of data [19]. The explosive growth in databases has created a need to develop technologies that use information and knowledge intelligently. Therefore, DMT has become an increasingly important research area [17]. Of the data mining techniques developed recently, several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining, are herein reviewed. The techniques for mining knowledge from different kinds of databases, including relational, transactional, object oriented, spatial and active databases, as well as global information systems, are also examined. Potential data mining applications and some research issues are discussed.

## 5. DATA MINING TECHNIQUES, KDD

### 5.1 Knowledge Discovery

The Knowledge discovery in databases (KDD) process is used to remove useful knowledge from volume data. The data mining refers to one meticulous step in this process. Data mining is the process of analyzing data from unlike perspectives and shortening it into useful information. Technically, data mining is the process of finding correlations or patterns amongst dozens of fields in large relational databases. Bellows show the different steps for extracting useful data from volume data [3].

### 5.2 Data –Mining Techniques

Data-mining techniques basically are pattern discovery algorithms. Some techniques such as association rules are only one of its kinds to data-mining, but most are drawn from related fields such as machine learning or pattern recognition. In this part we present the well-known data-mining techniques that have broadly used in intrusion detection.

## 6. KDD Cup'99 DATA SET

The proposed method is evaluated over the KDD Cup1999data. It contains a wide variety of intrusions replicated in military network environment. Each sample in the data 218 is a record of extracted features from a network connection gathered during the replicated intrusions. A connection is a sequence of TCP packets to and from various IP addresses. Connection documentation consists of 41 fields. It contains basic features about TCP connection as time protocol type, number of bytes transferred, domain specific type as number of file creation, number of unsuccessful login attempts, and whether root shell was obtained. It provides 100,000 labeled data items, composed of 99,999 normal samples and 1,000 attack samples [5].The KDD'99 was simulated in a military network environment and used for The Third worldwide Knowledge Discovery and Data Mining Tools opposition, which was held in conjunction with KDD-99. The contest task was to learn a projecting model or a classifier accomplished of distinguishing between legitimate and dishonest connections in a computer network. This data set contains one type of normal data and 24 different types of attacks that are characterized into four types such:

### 6.1 Denial of Service Attack (DOS)

A type of attack on a network that flood it with a waste of time traffics by the utilization of resources and memories.

### 6.2 Users to Root Attacku (U2R):

The attacker login a normal user relation on the system with target to get to access to the classification.

### 6.3 Remote To Local Attack (R2L):

Is when the attacker challenge to get a local access as a user of a piece of equipment on a network which he doesn't have any accurate to access to system.

### 6.4 Probe Attack

This attack is in relation to collecting in turn from a network of computers for a later use.

A generic view of the knowledge discovery process is outlined in, which consists of the following basic steps[23]:

- Developing and considerate the application domain, the related prior knowledge, and identifying the goal of the Knowledge Discovery in Databases process.
- Creating target data set.
- Data cleaning and preprocessing: necessary operations such as the elimination of noise, management missing data fields.
- Data decrease and projection:
  - Finding useful features to represent the data depending on the goal of the task.
  - Using dimensionality decrease or transformation methods to reduce the successful number of variables under kindness or to find invariant representation of data.
- Matching the goal of the Knowledge Discovery in Databases process to a particular data mining method.

## 7. LITERATURE REVIEW

**MOHAMMAD WAZID ET AL [2005]** Security is the biggest concern in Wireless Sensor Networks (WSNs) particularly for the ones which are categorized for military applications. They are prone to various attacks which degrades the network performance very fast. Sometimes multiple attacks are launched in the network using hybrid anomaly. In this condition it is very complicated to find out which kind of difference is activate. Proposed a hybrid anomaly detection technique with the request of k-means clustering. The study of the network data set consists of traffic data and end to end hindrance data is performed. The data set is clustered using weka 3.6.10. After clustering, we get the threshold values of different network presentation parameters (traffic and delay) [7]. These threshold values are used by the hybrid anomaly recognition technique to detect the anomaly.

**CHUNFU JAI ET AL [2009]** there are two technologies in intrusion detection systems: misuse detection and anomaly detection. Together misuse detection and anomaly detection contain advantages and disadvantages. At current, the intrusion detection system is residential by using these two technologies in conjunction with one a different, but there is not an effective method to evaluate the intrusion detection systems joint detection's performance. It is required to analyze it by establishing stringently mathematical equations. Considering the information assumption method to breakdown this problem, the intrusion detection capability can be used to analysis and evaluation. By dissimilarity two intrusion detection systems, it turns absent, the system that based on misuse and anomaly collaborative detection has the better detection effects [8].

**T.R GOPALKRISHNAN ET AL T.R [2011]** State of the art research in data mining is focusing on loosely distributed regionalized large scale databases use cloud computing for business application. Cloud computing poses a diversity of challenges in data mining operation arise out of the dynamic arrangement of data distribution as against the use of typical database scenarios in conventional architecture. Understanding of maximum efficiency depends much on the initiation of correct decision data mining [12].

**CHIRAG N.MODI ET AL [2012]** One of the major security issues in Cloud computing is to detect malicious actions at the network layer. In this, we propose a framework integrate network intrusion detection system (NIDS) in the Cloud. Our NIDS component consists of grunt and signature a priori algorithm. It generates new rules from captured packets. These new rules are appended in the Snort arrangement file to improve efficiency of Snort. It aims to detect identified attacks and derived of known attacks in Cloud by monitoring network traffic, while ensuring low false positive rate with practical computational cost [13].

**KAPIL WANKHADE ET AL [2013]** Intrusion Detection System (IDS) is an appropriate vital component of any network in today's world of Internet. IDS are a successful way to detect different kind of attacks in an inter related network there by secure the network. A successful Intrusion Detection System requires high correctness and detection rate as well as low false alarm rate. This focuses on a hybrid move on for intrusion detection system (IDS) based on data mining techniques. The main research process is clustering examination with the plan to improve the detection rate and reduce the fake alarm rate. Most of the earlier methods suffer from the disadvantage of k-means method with low detection rate and high false alarm rate [9].

**NUTAKARN MONGKONCHAI ET AL [2014]** Due to a fast growth of Internet, the number of network attacks has rise leading to the basics of network intrusion detection systems (IDS) to maintain the network. With mixed accesses and huge traffic volumes, several pattern identification techniques have been bringing into the research community. Data Mining is one of the analyses which many IDSs have adopt as an attack acknowledgment scheme. Thus, in this the classification methodology including characteristic and data selections was drained based on the well–known agreement schemes, i.e., Decision Tree, Ripper Rule, Neural Networks, Naïve Bays, *k*–Nearest–Neighbor, and Support Vector Machine, for intrusion detection analysis using both KDD CUP data set and recent HTTPBOTNET attacks[10].

## 8. CLUSTER ANALYSIS METHODS IN DATA MINING

Cluster analysis is a very important data mining technology to divide the data object into several meaningful subclasses, so that the members from the same clusters are quite similar and members from different clusters are quite different from each other. Therefore this method is applied for classifying log data and detecting intrusions. Clustering is an unsupervised knowledge method of data mining that takes unlabeled data points and tries to group them according to their similarity. In unsupervised approach there is no need of prior knowledge about training data whereas in supervised approach, given a set of normal data need to train in order to detect whether the test data belongs to normal or anomalous behavior. The general steps for clustering are: feature mining from sample data where input is sample data and output is matrix. Then implementation of clustering algorithm to access cluster genealogy diagram i.e. to reflect all the classification. After obtain a cluster genealogy diagram, the domain experts will decide the threshold selection according to the specific application by experience and domain knowledge. Data pre-processing, this includes standardization, integration; normalization etc. is one of the important step before applying Data mining[5]. This is also necessary precondition for normal operation of clustering. Clustering algorithm can be considered into four main groups: partitioning algorithm, hierarchical algorithm, density based algorithm and grid based algorithm.

### 8.1 Density-Based Methods

Most partitioning methods cluster objects based on the distance between objects. Such methods can find just spherical-shaped clusters and come across difficulty at discover clusters of arbitrary shapes. Other clustering methods have been residential based on the concept of density. The general design is to continue growing the given cluster as long as the density i.e. number of substance or data points in the neighborhood exceeds some threshold. It means that for each data point within a given cluster, the neighborhood of a given radius has to include at least a lowest amount number of points. Such a system can be used to filter out noise or outliers and decide clusters of arbitrary shape.

### 8.2 Grid –Based Methods

Grid-based methods divide the object space into a finite number of cells that form a grid configuration. All of the clustering operation is performed on the grid structure. The main advantage of this come near is its fast dispensation time, which is typically dependent mainly on the number of cells in each dimension in the quantized space.

### 8.3 Hierarchical Ciuster

Hierarchical Clustering methods are Agglomerative hierarchical methods. This Begins with as lots of clusters as substance. Clusters are successively merged until only one cluster mains. Divisive hierarchical methods start on with all objects in one cluster. Groups are continually divided until there are as many clusters as objects.

### 8.4 Partition Algorithm

Partitioning algorithm divides database of N objects into K clusters. Usually start with an
Initial separation and then use an iterative control strategy to optimize an objective function. Classification of clustering algorithm.

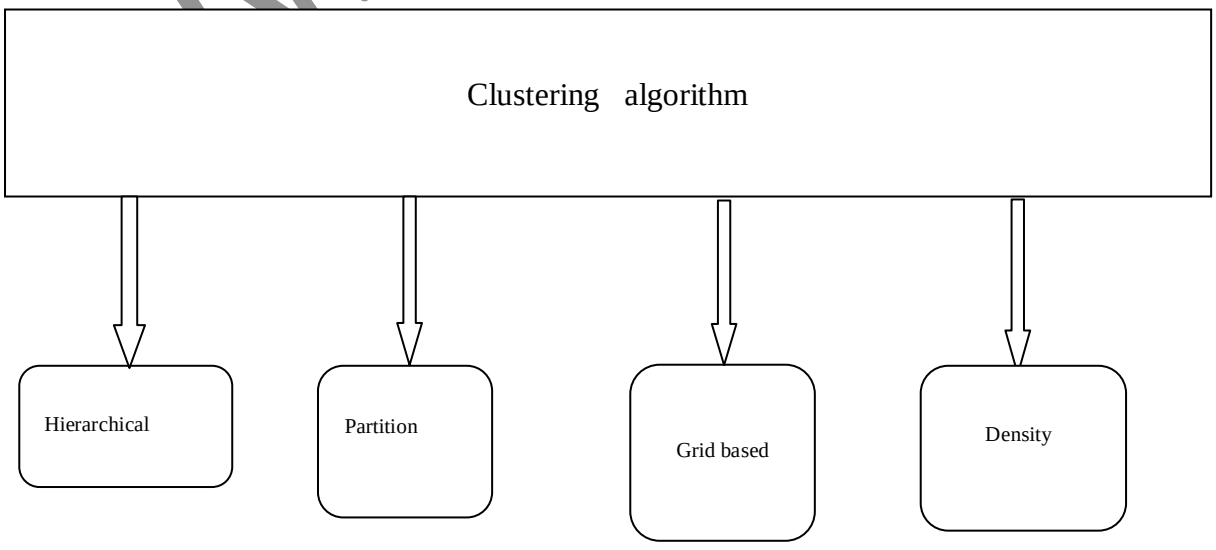


**Fig. 8.1 Classification of Clustering Algorithm [5]**

## 9. K-MEANS CLUSTERING

K-means clustering algorithm is one of the commonly used Partition based clustering algorithm. It is censored and iterative based clustering by low time complexity and fast convergence which is very important in intrusion detection due to large size of network traffic audit dataset .K-means algorithm divides N data object into K clusters. The objects in the same clustering have higher similarity while objects in different clustering have smaller similarity. It is a dynamic clustering based on standard measure function. K-means algorithm divides N vectors into K classes. Usually start with an initial partition then use an iterative control strategy to optimize an objective function [3]. K-means represents a type of useful clustering techniques by competitive learning which is also one of the promising techniques in intrusion detection [11]. The algorithm has following steps:

- We place k points into the breathing space represented by the objects that are being clustered. Initial group cancroids are represented by these points.
- Assign each entity to the group that has the closest centurions.
- After the assignment of all objects, recalculate the positions of the k cancroids.
- Repeat Steps 2 and 3 until the cancroids no longer move [7]

**TABLE-9.1 Comparison Between Many Techniques**

| S. No | Authors | Advantage | Techniques | Methods | Limitations |
|---|---|---|---|---|---|
| **1** | S.V.Shirbhate | Discovered unknown attack | Hybrid anomaly detection technique[5] | K-means clustering. | Require more cluster |
| **2** | A.M. Chandrasekhar | High Accuracy | Support Vector Machine(SVM)[20] | Classification | The training and testing speed is show |
| **3** | Yang Yong | It solves the problems with multiple solutions | Genetic Algorithms[19] | REGAL System | No constant optir mizati On response Time. |
| **4** | Subaira.A.S | Highly tolerate the noisy data. | Neural Networks[4] | MLP(Multilayer perceptron) | It Requires long Training time. |
| **5** | G. J. Klir | 5.Rule base or fuzzy sets easily modified | Fuzzy Logic[22] | Classification | Hard to develop a Model from a fuzzy System. |
| **6** | Niken Prasasti | Can handle high dimensional Data. | Decision Tree[21] | Classification | Limited to one output Attribute. |

## CONCLUSION

In this paper functions, securities, techniques and problems of ids are discussed. Ids current industrial intrusion detection systems make use of misuse detection. As such, they completely are short of the ability to detect new attacks. It is impossible to prevent security violation completely by using the exciting security technology. Accordingly, Intrusion. Afterwards we introduced the data-mining and machine learning algorithms and it advantages of the IDS based on data mining and machine learning approach such as clustering and classification.

## REFERENCES

[1] Kalpana Jaswal, Seema Rawat,Praveen Kumar "Design and Development of a prototype Application for Intrusion Detection using Data mining" 2015 IEEE.
[2] G.V. Nadiammai, " Effective approach toward Intrusion Detection System using data mining techniques"2013.
[3] Nadya EL Moussaid, Ahmed Toumanari Essi, "Overview of Intrusion Detection Using Data-Mining and the features selection"IEEE 2015.
[4] Subaira.A.S, Mrs. Anitha.P " Efficient Classification Mechanism for Network Intrusion Detection System Basedon Data Mining Techniques" 8th Proceedings International Conference on Intelligent Systems and Control (ISCO) IEEE2014.
[5] S.V Shirbhate, Dr.S.S.Sherkar ,Dr.V.M.Thakare "Performance Evaluation of PCA Filter In Clustered Based Intrusion DetectionSystem" . 2014 International Conference on Electronic Systems, Signal Processing and Computing IEEE 2014.
[6] Mr. Mohit Sharma, Mr. Nimish Unde, Mr. Ketan Borude, A Data Mining Based Approach towards Detection of Low Rate DoS Attack"2014. International Conference for Convergence of Technology – IEEE 2014.
[7] Mohammad Wazid " Hybrid Anomaly Detection using K-Means Clustering in Wireless Sensor Networks"IEEE 2005.
[8] Chunfu Jia" Performance Evaluation of a Collaborative Intrusion Detection System"IEEE 2009.
[9] Kapil Wankhade, Sadia Patka " An Efficient Approach for Intrusion Detection Using Data Mining Methods"IEEE 2013.
[10] Chakchai So–In, Nutakarn Mongkonchai, Phet Aimtongkham. "An Evaluation of Data Mining Classification Models for Network Intrusion Detection"IEEE 2014.
[11] Li Hanguang, Ni Yu "Intrusion Detection Technology Research Based on Aprior Algorithm" 2012
[12] T.R. Gopalakrishnan Nair, K.Lakshmi Madhuri" Data Mining USsing Hierachical Virtual K-Means Apporaoch Intergrating Data Fragments In Cloud Computing Envioroment "IEEE 2011.
[13] Muttukrishnan Rajaraja " Integrating Signature Apriori based Network Intrusion DetectionSystem (NIDS) in Cloud Computing"2012.
[14] Ketan Sanjay Desale, Chandrakant Namdev Kumathekar, Arjun Pramod Chavan "Efficient Intrusion Detection System using Stream Data Mining Classification Technique "2015.International Conference on Computing Communication Control and Automation.
[15] Muamer N. Mohammada, Norrozila Sulaimana, Osama Abdulkarim Muhsinb "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment"2010.
[16] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan "HsiaoData mining techniques and applications – A decade review from 2000 to 2011"E IEEE 2010.
[17] Fayyad, U., Djorgovski, S. G., & Weir, N."Automating the analysis and cataloging of sky surveys".
[18] Ha, S., Bae, S., & Park, S. (2000)." Web mining for distance education. In IEEE international conference on management of innovation and technology"IEEE 1998.
[19] Yang Yong" The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm" 2011.
[20] A.M. Chandrasekhar,K.Raguveer"Intrusion Dictation using k-means,fuzzy neural network and svm classifiers"2013.
[21] Niken Prasasti,Hayato Ohwada" Applicability of Machine-Learning Techniques in Predicting Customer Defection" 2014.
[22] G. J. Klir, "Fuzzy arithmetic with requisite constraints", Fuzzy Sets and Systems, 1997.
[23] Charles A. Fowler and Robert J. Hammell "Converting PCAPs into Weka Mineable Data"IEEE 2014.